Application of two-dimensional infrared spectroscopy to benchmark models for the amide I band of proteins

Bondarenko, Anna S.; Jansen, Thomas L. C.

Published in:
Journal of Chemical Physics

DOI:
10.1063/1.4919716

Publication date:
2015

Application of two-dimensional infrared spectroscopy to benchmark models for the amide I band of proteins

Anna S. Bondarenko, and Thomas L. C. Jansen

Citation: The Journal of Chemical Physics **142**, 212437 (2015); doi: 10.1063/1.4919716
View online: https://doi.org/10.1063/1.4919716
View Table of Contents: http://aip.scitation.org/toc/jcp/142/21
Published by the American Institute of Physics

Application of two-dimensional infrared spectroscopy to benchmark models for the amide I band of proteins

Anna S. Bondarenko and Thomas L. C. Jansen
Zernike Institute for Advanced Materials, University of Groningen, Nijenborgh 4, 9747 AG Groningen, The Netherlands

(Received 9 February 2015; accepted 23 April 2015; published online 7 May 2015)

In this paper, we present a novel benchmarking method for validating the modelling of vibrational spectra for the amide I region of proteins. We use the linear absorption spectra and two-dimensional infrared spectra of four experimentally well-studied proteins as a reference and test nine combinations of molecular dynamics force fields, vibrational frequency mappings, and coupling models. We find that two-dimensional infrared spectra provide a much stronger test of the models than linear absorption does. The best modelling approach in the present study still leaves significant room for future improvement. The presented benchmarking scheme, thus, provides a way of validating future protocols for modelling the amide I band in proteins. © *2015 AIP Publishing LLC.* [http://dx.doi.org/10.1063/1.4919716]

I. INTRODUCTION

In recent years, two-dimensional infrared (2DIR) spectroscopy,[1] which is more sensitive to couplings and dynamics than conventional linear absorption spectroscopy,[2] has been developed to probe structure and dynamics of protein systems by probing the amide I (CO stretch) vibrations.[1,3–26] Each amino acid unit in a protein has one CO stretch vibration in the backbone chain[27,28] and some side chains absorb in the same region.[29,26] Furthermore, the transition dipoles of the CO vibrations are large resulting in delocalised vibrational modes arising from the coupling between different CO oscillators.[3,17,30] This results in very congested spectra that are very challenging to interpret. Theoretical approaches have been developed to help unravel the structure and dynamics determining the spectral line shapes.[30–34] Still, very few tests have been provided for these simulation methods. The goal of this paper is to provide a new benchmark test for computational models employed for the amide I region of proteins by comparing the simulated linear absorption and 2DIR spectra of selected proteins with well-known structure to the experimental data. In this way, a measure is devised directly for testing of simulation methods under the conditions that one ultimately want to be able to calculate.

The simulation of 2DIR spectra of the amide I band of proteins consists of several components. First, the dynamics of the protein must be modelled using molecular dynamics (MD) simulations, which require the use of a classical force field. Then, the vibrational Hamiltonian must be constructed typically employing mappings allowing to extract local mode vibrational frequencies and couplings from the structural information of the MD simulations. Finally, this information must be converted to a two-dimensional spectrum typically using response function approaches derived using time-dependent perturbation theory.[2,35]

Numerous force fields have been developed to model the atomistic structure of proteins.[36–39] A number of extensive benchmarking studies of force fields using knowledge from nuclear magnetic resonance (NMR) spectroscopy or x-ray diffraction have been presented.[40–43] These studies show that different force fields predict quite different stability of the various secondary structural elements. From one of the most recent studies using extensive NMR data,[41] the AMBER 99SB force field with improved side chain torsion potentials (AMBER 99SB-ILDN) came out the best.[37] We will therefore include this particular force field in the present study.

Numerous models for the amide I frequency have been presented during the last decades.[44–69] Most of these models attribute the solvent shift to the dominant contribution of electrostatic interactions. However, recent studies have suggested important contributions from dispersion, charge transfer, and polarisation effects.[64,66,67,70–73] The electrostatic frequency mappings have been tested against *ab initio* calculations,[50,51,69,74] solvent experiments of single amide units,[52,60,73,75] gas phase experiments[76,77] of small peptides, and linear absorption of proteins.[62,64] So far, only one mapping for the amide unit preceding proline has been presented.[59] This amide unit is special as it is a tertiary amide group. The primary amide groups present in asparagine and glutamine have also only been considered in a single study.[60] The typical mappings have not taken the special conditions of terminal groups into account with the most prominent exception of a study, where an empirical map was developed using dipeptides, thus, in reality only including terminal groups.[63] Finally, a number of amino acids have other vibrations in the vicinity of the amide I region.[26,29] These vibrations are mostly ignored in studies of the amide I region.

Different models have been proposed for the short (nearest neighbour)[46,50,53,59,74,78–82] and long[74,83,84] range couplings. The couplings have been tested on the linear spectra of

proteins,[62,64,84] using isotope edited 2DIR,[85,86] *ab initio* calculations,[50,51,74] and gas phase experiments.[76,77] In general, benchmarking the coupling models is the most challenging as the observables depending on the couplings inherently also depend on the vibrational frequencies.

2DIR spectroscopy resembles two-dimensional nuclear magnetic resonance spectroscopy[87] in many ways; however, the technique relies on the use of infrared laser pulses. Four coherent laser pulses are applied to the sample separated by three time delays. The first and the last of these time delays are denoted the coherence times and labeled t_1 and t_3. The time delay in the middle is denoted the waiting time, t_2. Typically, the signal is Fourier transformed for the coherence times to provide an excitation frequency ω_1 and a detection frequency ω_3. Different experimental setups exist with variation of the exact details, of which a description is beyond the discussion of the present paper.[2] The important advantage of the method is that it allows correlating frequencies of vibrations present in the sample before and after the waiting time. This makes 2DIR a powerful tool that allows studying couplings between different vibrations,[1,30,3,88] chemical exchange processes,[89,91,90,92–94] and spectral diffusion processes.[95–98]

Here, we will present a novel benchmarking approach utilising the 2DIR spectra of four proteins with well-known structure. This will provide both a way to select the best combination of force field, frequency mapping, and coupling model and a critical test of the state-of-the-art models. Considering the vast number of available models, it should be recognised that it is beyond the scope of the present study to present an exhaustive test of all possible combinations of modelling parameters. The outline of the remainder of this papers is as follows. In Sec. II, we will describe our simulation protocol and the benchmarking approach. In Sec. III, we present the benchmarking results using linear absorption and 2DIR spectroscopy utilising four different proteins to test nine different combinations of force field, electrostatic mapping, and coupling model. Finally, we will draw the conclusions in Sec. IV.

II. METHODS

The set of proteins used in this study was selected to include a variety of secondary structures ranging from proteins predominantly containing α-helices to proteins dominated by β-sheet content. Our choice was restricted to proteins for which experimental FTIR and 2DIR data are readily available[17,21] and proteins with well determined crystal structures. These restrictions lead us to choose the following four proteins: myoglobin (Mb, 153 residues, Protein Data Bank Identification (PDB ID) 5MBN;[99] 0% β-sheet, 89% α-helix structure, 11% coils) that has no β-sheets; lysozyme (Lys, 129 residues, PDB ID 1AKI;[100] 10% β-sheet, 52% α-helix structure, 38% coils) that has a small three-stranded β-sheet region; ribonuclease A (RNse A, 124 residues, PDB ID 1FS3;[101] 40% β-sheet, 26% α-helix structure, 34% coils) that has two domains that vary from two- to four-stranded regions; and concanavalin A (Con A, 237 residues, PDB ID 1NLS;[102] 55% β-sheet, 8% α-helix structure, 37% coils), the most extended system, with two relatively flat six-stranded anti-parallel β-sheets. The chosen proteins are illustrated in Fig. 1.

MD simulations were performed with GROMACS-4.6.5[103] using GROMOS 54a7[104] and Amber 99SB-ILDN[37] force fields. The last force field was used for all proteins

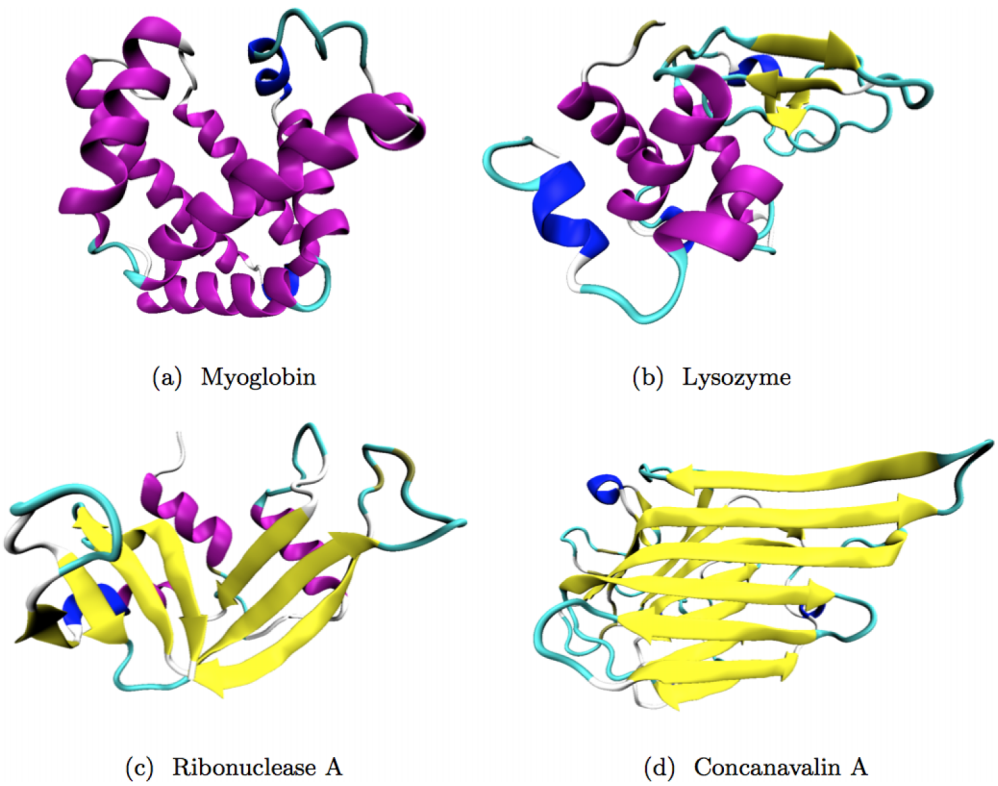

FIG. 1. The structures of the selected proteins, visualised with Visual Molecular Dynamics (VMD).[118] The secondary structure is highlighted with the following colours: α-helical regions are magenta, β-sheet regions are yellow, 3^{10}-helical regions are blue, and random coil regions are cyan.

except myoglobin that contains a heme group for which no parameters are implemented in this Amber force field. The protein crystallographic structures were downloaded from the Protein Data Bank.[99–102] Missing protons were added, each protein was placed in a cubic box with at least 1.0 nm between the protein surface and the box edge, and solvated in explicit water solvent SPC/E (extended simple point charge) water[105] (between 9480 and 17 700 water molecules depending on the size of the protein). Chlorine or sodium counter-ions were added to keep each simulation box neutral. Finally, the systems were relaxed by energy minimisation. Equilibration was conducted in two phases: 100-ps NVT equilibration[106] that stabilised the temperature of the system and 100-ps NPT equilibration[106] that stabilised the pressure (and, thus, also the density) of the system. Afterwards, a 1-ns MD simulation with 2 fs time step was performed, using the NPT ensemble with Nose-Hoover thermostat[107] and the Parrinello-Rahman barostat.[108] Long range electrostatic interactions were calculated with the PME (Particle Mesh Ewald)[109] algorithm using a cutoff of 1 nm for the real space treatment. The atomic positions along the trajectories were stored for every 20 fs, giving a total of 50 000 stored snapshots for each simulation. To test the flexibility of the structures, the root mean square fluctuations of the concanavalin A structure were calculated and compared with that found for a longer 10-ns trajectory (see Figure S2 of the supplementary material[110]). Apart from two very short regions, the root mean square fluctuations of the structures are the same suggesting that using the 1-ns trajectories is a reasonable approximation covering the majority of the flexibility of the protein structures.

For every snapshot of MD trajectory, an amide I Hamiltonian of the following floating oscillator form was constructed:

$$H(t) = \sum_i \omega_i(t) B_i^\dagger B_i - \sum_i \frac{\Delta}{2} B_i^\dagger B_i^\dagger B_i B_i + \sum_{i,j} J_{ij}(t) B_j^\dagger B_i + \sum_i \vec{\mu}_i(t) \cdot \vec{E}(t)(B_i^\dagger + B_i). \quad (1)$$

Here, $B_i^\dagger$ and B_i are the bosonic creation and annihilation operators. The indices i and j number the local amide I vibrations along the backbone. The site frequencies are denoted ω_i and are time-dependent as the values are determined by either of the electrostatic mappings here denoted the Jansen map,[52] the Skinner map,[60] and the Tokmakoff map.[63] The Tokmakoff and Skinner maps make use of a linear correlation of the electric field on atoms in each amide group with the frequency of the amide I vibration. The correlation coefficients were determined empirically by fitting using the electric fields generated by the CHARMM-27 (Chemistry at Harvard Macromolecular Mechanics, nr. 27) and GROMOS-53a6 force fields for the Tokmakoff and Skinner maps, respectively. Here, we used the electric fields as generated by the force fields used in the MD simulations. The Jansen map includes both electric fields and their gradients on the atoms in the amide group in a linear correlation with the amide I frequency. In this case, the coefficients were determined from density functional theory (DFT) calculation of the N-methyl acetamide (NMA) molecule in different point charge environments. In all cases, the tertiary amide units preceding proline were treated with the only existing electrostatic map for this group,[59] which essentially employs the Jansen mapping strategy. In all cases the frequency shifts from the nearest neighbour amide groups along the backbone were treated with the Ramachandran angle based map described in Ref. 74. In all treatments, only the backbone amide groups were included in the Hamiltonian. The anharmonicities, Δ, were fixed to 16 cm^{-1}.[1] The transition dipoles, $\vec{\mu}_i$, were taken from the mapping in Refs. 52 and 82, when the Jansen or Tokmakoff mappings were used, and from Ref. 79, when the Skinner mapping was used. The external laser field used to excite the vibrations is denoted $E(t)$. The couplings, J_{ij}, were divided into two types. The coupling between nearest neighbours was always treated with the Ramachandran angle based mapping derived from DFT calculations on glycine dipeptide,[74,82] with the only exception for couplings involving tertiary amides, where a similar mapping based on dipeptides of glycine and proline was employed.[59] Long range couplings, i.e., all other couplings, were treated with either a transition charge coupling (TCC) scheme[74,80] or a transition dipole coupling (TDC) scheme.[79,111] The TCCs are given by the equation

$$J_{ij} = \frac{1}{4\pi\epsilon_0} \sum_{n,m} \left(\frac{dq_n dq_m}{|\vec{r}_{n_i m_j}|} - \frac{3 q_n q_m (\vec{v}_{n_i} \cdot \vec{r}_{n_i m_j})(\vec{v}_{m_j} \cdot \vec{r}_{n_i m_j})}{|\vec{r}_{n_i m_j}|^5} - \frac{dq_n q_m \vec{v}_{m_j} \cdot \vec{r}_{n_i m_j} + q_n dq_m \vec{v}_{n_i} \cdot \vec{r}_{n_i m_j} - q_n q_m \vec{v}_{n_i} \cdot \vec{v}_{m_j}}{|\vec{r}_{n_i m_j}|^3} \right). \quad (2)$$

Here, the atomic charges, q_n, transition charges, dq_n, and normal mode coordinates, $\vec{v}_{m_i}$, are taken from Ref. 74. The index n runs over the atoms in amide unit i and the index m runs over atoms in amide unit j. The symbol, $\vec{r}_{n_i m_j}$, represents distance vectors between atoms. The TDCs are given by

$$J_{ij} = \frac{1}{4\pi\epsilon_0} \left(\frac{\vec{\mu}_i \cdot \vec{\mu}_j}{r_{ij}^3} - 3 \frac{(\vec{\mu}_i \cdot \vec{r}_{ij})(\vec{\mu}_j \cdot \vec{r}_{ij})}{r_{ij}^5} \right). \quad (3)$$

Here, the transition dipoles, $\vec{\mu}_i$, were taken from Ref. 79 and the distances, $\vec{r}_{ij}$, are between the transition dipole positions of the two carbonyls, located at the position $r_c + 0.665$ Å $\cdot\, n_{CO} + 0.258$ Å $\cdot\, n_{CN}$, where r_c is the position of the carbonyl carbon and n_{CO} and n_{CN} are unit vectors along the CO and CN bonds, respectively, as described in Ref. 79.

The linear absorption and 2DIR spectra were calculated using the Numerical Integration of the Schrödinger Equation (NISE) method,[32,33] which is based on numerically solving the time-dependent Schrödinger equation for the amide I Hamiltonian to determine the linear and third-order response functions governing the linear absorption and 2DIR spectroscopies,[2,32] respectively. The linear response functions were calculated for 500 equally spaced starting points along the Hamiltonian trajectories and summed up. These linear response functions

were calculated for times from zero to 5.12 ps and Fourier transformed to obtain the spectra in frequency domain. For the 2DIR simulation, the two coherence times were varied from zero to 5.12 ps as well. In the present simulations, the waiting time was kept fixed at 0, which is the value for which the experiments were performed. The 2DIR response functions were calculated for every 2 ps along the trajectory resulting in 500 realisations that were summed up to give the total response functions that were then Fourier transformed to give the 2DIR spectra. The response functions were simulated for the 21 unique molecular dynamics frame polarization directions[112] and averaged to obtain the 2DIR spectra corresponding to the experimentally employed laser configuration, where the polarization of the two first laser pulses is perpendicular to those used for the detection.[112] We utilised a vibrational lifetime of 1 ps during the coherence times leading to a slight homogeneous broadening of the spectra.[2]

III. RESULTS

A. Linear spectra

Experimental FTIR spectra of the four proteins[21] are shown in Fig. 2(a). Qualitatively, absorption spectra associated with primarily the α-helical proteins myoglobin and lysozyme are characterized by a single band centred near 1650 cm^{-1} with an approximate diagonal width of 50 cm^{-1}, while proteins composed of primarily β-sheet show a peak centred near 1620-1640 and a shoulder at 1680 cm^{-1}, respectively, resulting from vibrations whose main transition dipoles lie perpendicular and parallel to the β-strands, respectively.[17,79] The experimental sample of concanavalin A tends to aggregate or oligomerize during the measurement. At low concentration (~4 mg/ml), the FTIR spectrum has a flat peak,[21] and at high concentration (~15 mg/ml), a low frequency peak grows in resulting in a more triangular shaped peak.[17] We used the most recent experimental data for concanavalin A corresponding to the low concentration situation.[21] From the FTIR spectra, it is evident that going from exclusively α-helix system of myoglobin to exclusively β-sheet system of concanavalin A, the peak amide I transition frequency red-shifts from 1650 cm^{-1}, the absorption lines change their shape from symmetric to more skewed towards the blue, and the peaks are getting broader.[113] This trend is expected with increasing β-sheet content. Nevertheless, it is difficult to correlate the spectral profiles observed in FTIR spectra with different structural elements present in this series of four proteins.[21] In the FTIR spectra, the two-peak structure indicating the presence of a β-sheet is hardly evident as the high frequency peak is just seen as a weak shoulder for concanavalin A and ribonuclease A, while no such shoulder is present for myoglobin and lysozyme.

We used nine different models for our benchmark study. These nine models were composed by varying three different types of parameters. We used two different MD force fields, namely, GROMOS 54a7[104] and Amber 99SB-ILDN.[37] We used two different models for the long range coupling: TCC[74] and TDC.[79] And finally, we used three different electrostatic maps for the amide I site frequency fluctuations: the Jansen map,[52] the Skinner map,[60] and the Tokmakoff map.[63] The Amber force field was only used together with the TCC long range coupling. All other combinations of parameters were tested. Both FTIR and 2DIR spectra were calculated with the nine resulting parameter combinations. In the supplementary material,[110] an extensive overview of the FTIR data is provided (Figures S3-S6).

To analyse the FTIR spectra, we shifted the peak position to maximize the spectral overlap between the simulated and experimental spectra as defined in Eq. (4) below. This shift, $\Delta\omega$, was determined for each protein and is presented in Table I and Fig. 3. The Jansen mapping scheme systematically overestimates the frequencies, while the Skinner and Tokmakoff mappings underestimate the peak frequencies. The range of peak position shifts for each protein is around 30 cm^{-1}. A general trend in the error of the predicted peak frequencies is observed, when going from α-helical to β-sheet dominated structures. The blue-shift observed for the Jansen map simulations is increasing when going from myoglobin to concanavalin A, while for the Skinner and Tokmakoff maps the red-shift is increasing, when going from concanavalin A to myoglobin. If the peak shifts are systematic, one can compensate for the error in the predicted frequency by shifting the overall spectrum by a fixed value. The standard deviations of the errors in the peak positions are presented in Table I and generally smaller for the

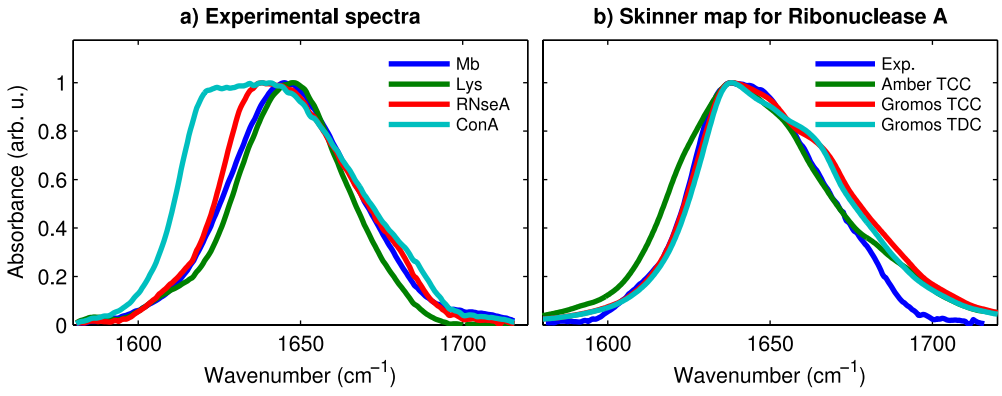

FIG. 2. FTIR spectra: (a) experimental spectra of the four proteins; (b) predicted FTIR spectra of ribonuclease A, using Skinner map, and different force fields and coupling maps.

TABLE I. The values of the frequency shifts in cm^{-1} needed to maximize the spectral overlap in the FTIR spectra as defined by Eq. (4).

		Mb	Lys	RNse A	Con A	Average	Standard deviation
Jansen	Amber, TCC	...	−15.7	−15.6	−20.3	−17.2	2.2
	Gromos, TCC	−9.2	−17.3	−19.6	−26.1	−18.0	6.0
	Gromos, TDC	−10.9	−18.8	−21.5	−26.8	−19.5	5.7
Skinner	Amber, TCC	...	15.2	13.8	11.1	13.4	1.7
	Gromos, TCC	18.8	10.3	6.4	2.2	9.4	6.1
	Gromos, TDC	16.9	8.5	5.1	0.9	7.8	5.9
Tokmakoff	Amber, TCC	...	11.6	13.1	9.5	11.4	1.5
	Gromos, TCC	27.5	19.3	15.3	6.7	17.2	7.5
	Gromos, TDC	−6.0	−9.9	−8.9	−11.3	−9.0	1.9

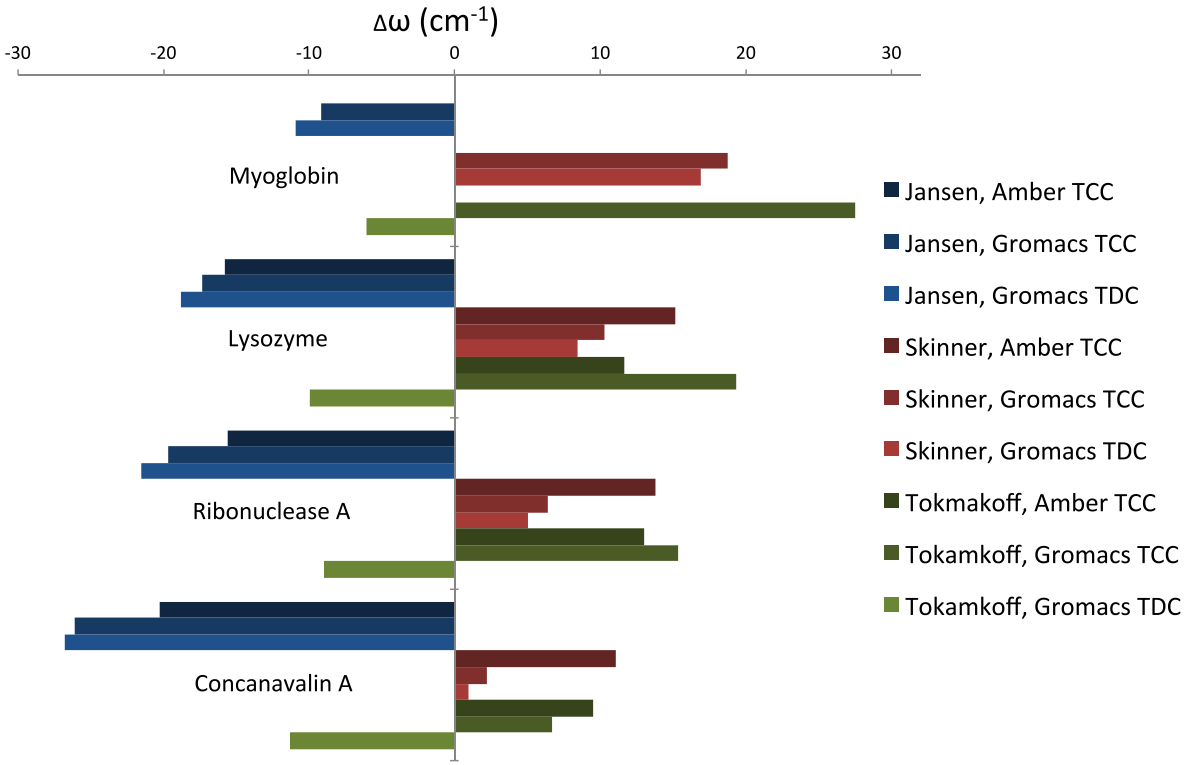

FIG. 3. Diagram of the frequency shifts needed to maximize the spectral overlap according to Eq. (4).

Jansen map than for the other mappings, suggesting that the errors in the peak positions are slightly easier to compensate, when using the Jansen map. In Figure S1 of the supplementary material,[110] the FTIR spectra are decomposed according to the secondary structure content. It is seen that the main peaks from α-helices and β-sheets are in close proximity of each other, where one would expect the peaks from α-helix structures to be higher in frequency.[28] This can explain the overall trend in $\Delta\omega$ observed when going from the α-helix rich myoglobin to the β-sheet rich concanavalin A.

To measure the similarity of the spectral line shapes quantitatively, the value of the spectral overlap, S^{1D}, between each simulated spectrum and its experimental counterpart was calculated as[114]

$$S^{1D} = \sum_i \left(I(\omega_i) I_{\mathrm{ref}}(\omega_i)\right) / \sqrt{\left(\sum_i I(\omega_i) I(\omega_i)\right) \times \left(\sum_i I_{\mathrm{ref}}(\omega_i) I_{\mathrm{ref}}(\omega_i)\right)}. \quad (4)$$

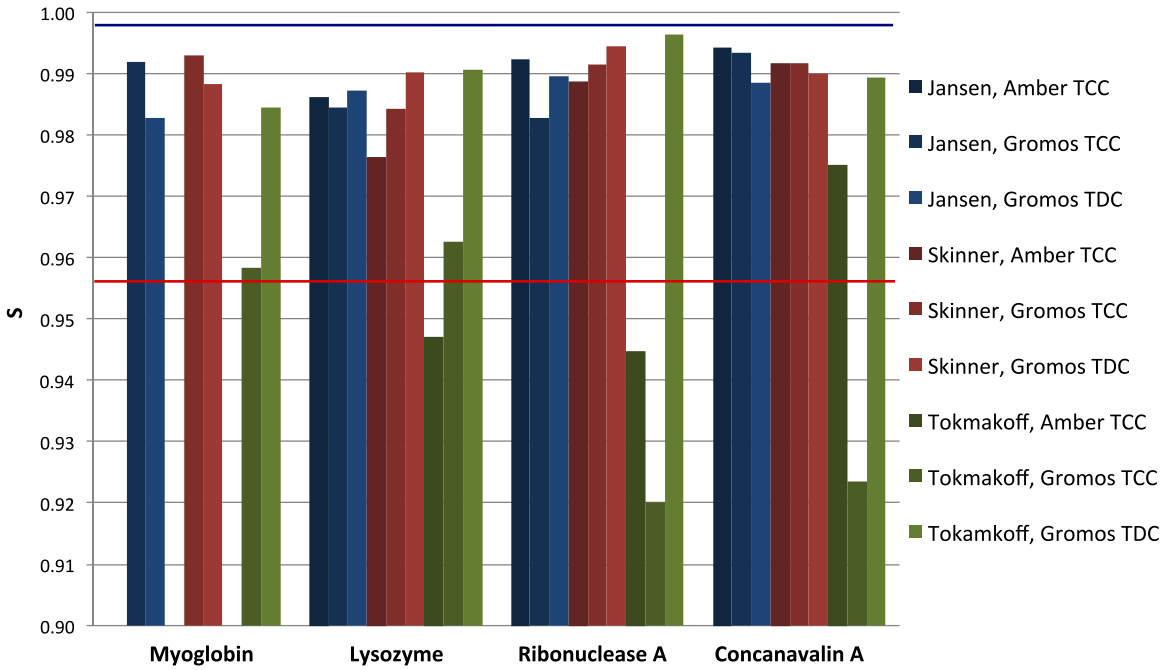

FIG. 4. Diagram of the spectra overlap between simulated FTIR spectra and experimental ones for each of the four proteins. Upper limit (blue line) is the highest spectral overlap between experimental spectra (myoglobin and lysozyme). Lower limit (red line) is the lowest spectral overlap between experimental spectra (lysozyme and concanavalin A).

Here, $I(\omega_i)$ is the intensity of the simulated spectra for the frequency ω_i, where the frequencies were shifted to maximize the overlap with the experiment. (Equivalent results using a shift to match the peak position are given in the supplementary material.[110]) $I_{\text{ref}}(\omega_i)$ is the experimental intensity at the same frequency. In general, the intensities are not known for the same frequencies for theory and experiment, and a cubic spline interpolation[115] was used to determine the intensity of the simulated spectra at the frequencies for which the experimental data were known. We note that this spectral overlap is equivalent to calculating the root mean square deviation between the signals using a normalization factor of $\sqrt{\sum_i I(\omega_i)^2}$ for each involved signal. For reference, the spectral overlap was calculated in an analogous manner between the experimental spectra of the different proteins. The significance of the overlap is that when the spectral shapes are identical, the overlap is one, and as the shapes get more and more different, this value drops towards zero.

The determined overlaps are presented in the diagram in Fig. 4 and Table II. For judging the quality of the models, two limits were used, presented by blue and red horizontal lines. The lowest value of 0.956 is the value of overlapping between the two proteins with the smallest overlap between their spectra as determined from the overlap of the experimental data (lysozyme and concanavalin A); therefore, in this limit, theoretical models are not able to differentiate between the two proteins that are most different. Essentially, models not able to provide an overlap above this value are useless. The highest value of 0.998 is the largest value of the overlap between the experimental spectra of the proteins (myoglobin and ribonuclease A). Ideally, one would like to be able to obtain overlaps with experiment above this value to be able to clearly distinct the different proteins from each other from simulation. As can be seen from the diagram, the best values are obtained with the Skinner map, where the overlap in all the cases is above the lower limit, and in the most cases close to the upper

TABLE II. The values of the spectral overlap between the shifted calculated FTIR spectra and experimental data.

		Mb	Lys	RNse A	Con A	Average	Standard deviation
Jansen	Amber, TCC	...	0.986	0.992	0.994	0.991	0.003
	Gromos, TCC	0.992	0.985	0.983	0.993	0.988	0.005
	Gromos, TDC	0.983	0.987	0.990	0.989	0.987	0.003
Skinner	Amber, TCC	...	0.976	0.989	0.992	0.986	0.007
	Gromos, TCC	0.993	0.984	0.991	0.992	0.990	0.003
	Gromos, TDC	0.988	0.990	0.994	0.990	0.991	0.002
Tokmakoff	Amber, TCC	...	0.947	0.945	0.975	0.956	0.014
	Gromos, TCC	0.958	0.963	0.920	0.924	0.941	0.019
	Gromos, TDC	0.985	0.991	0.996	0.989	0.990	0.004

TABLE III. The values of the spectral overlap between the shifted calculated 2DIR spectra and experimental data.

		Mb	Lys	RNse A	Con A	Average	Standard deviation
Jansen	Amber, TCC	...	0.815	0.851	0.805	0.823	0.020
	Gromos, TCC	0.736	0.826	0.843	0.716	0.780	0.055
	Gromos, TDC	0.731	0.811	0.817	0.669	0.757	0.061
Skinner	Amber, TCC	...	0.867	0.896	0.824	0.862	0.030
	Skinner, TCC	0.774	0.864	0.901	0.789	0.832	0.052
	Skinner, TDC	0.723	0.850	0.883	0.775	0.808	0.063
Tokmakoff	Amber, TCC	...	0.769	0.778	0.794	0.780	0.011
	Gromos, TCC	0.713	0.731	0.726	0.656	0.707	0.030
	Gromos, TDC	0.718	0.850	0.863	0.564	0.749	0.121

limit, and for ribonuclease A and concanavalin A it is above this limit. For the Jansen map, with the exception of two sets of parameters for lysozyme, all overlaps are above the lower limit. The worst values are obtained for the Tokmakoff map with a large number of values below the lower limit. One should note that all predicted overlaps are above 0.9 indicating that the FTIR spectra are generally quite similar.

The Tokmakoff map gives narrower spectra in comparison with the experimental spectra. For the β-sheet structures ribonuclease A and concanavalin A, the spectra even exhibit multiple peaks, likely resulting from too narrow distributions of the site frequencies. The Jansen map gives better results, reproducing the peak shape quite well, while the best results are given by Skinner map. In Fig. 2(b), spectra for ribonuclease A, where the spectra were most accurately predicted, are presented for the three different Skinner map parameter combinations. The best combination of parameters is Gromos force field with TDC scheme, where the left shoulder coincides with the experimental one, while the right shoulder is slightly wider.

B. 2DIR spectra

2DIR spectra are much more sensitive to couplings and the nature of frequency fluctuations than FTIR. Therefore, one may already expect that 2DIR spectroscopy is much more sensitive to structural differences in proteins than FTIR. We, thus, expect that 2DIR spectra are more discriminative for benchmarking than FTIR. To evaluate the quality of the nine simulation protocols with 2DIR, the spectral overlap between 2DIR spectra was determined using the following equation:[114]

$$S^{2D} = \sum_{i,j} \left(I(\omega_i,\omega_j) I_{\mathrm{ref}}(\omega_i,\omega_j) \right) / \sqrt{\left(\sum_{i,j} I(\omega_i,\omega_j) I(\omega_i,\omega_j) \right) \times \left(\sum_{i,j} I_{\mathrm{ref}}(\omega_i,\omega_j) I_{\mathrm{ref}}(\omega_i,\omega_j) \right)}. \quad (5)$$

Here, $I(\omega_i,\omega_j)$ is the two-dimensional spectral intensity for $\omega_1 = \omega_i$ and $\omega_3 = \omega_j$, while $I_{\mathrm{ref}}(\omega_i,\omega_j)$ is the corresponding experimental reference spectrum. As for the linear absorption case, a value of the spectral overlap of one corresponds to a perfect match and lower values indicate a worse agreement between the two spectra considered. Again to evaluate this quantity, the simulated spectral intensities must be known at the same points as for the experimental data. Therefore, the intensities of the simulated spectra were interpolated at the experimental frequency points, using bilinear interpolation in Matlab R2013a[115] after shifting the peak position of each spectrum identical to that used for the FTIR spectra to maximize the overlap of the FTIR spectra. Again, the spectral overlaps were also determined between the experimental spectra of the four proteins for reference.

The determined spectral overlaps are summarised in Table III and visualized in Fig. 5. The values of overlap are significantly lower than what were found for the FTIR spectra. The overlaps between the experimental spectra are lower than for linear spectra as the most different protein spectra have an overlap of just 0.792 (lysozyme and concanavalin A), while for the most similar spectra, the overlap is 0.967 (between myoglobin and ribonuclease A). Ideally, a simulation protocol should, thus, provide overlaps above this value and it would be essentially useless if values are below the lower value. Generally, the overlaps for the different simulations are between these two limits. All values for the Skinner map are higher than the lowest limit. The worst results for all maps are observed for concanavalin A. The spectral match is the best for ribonuclease A, where all simulations give overlaps above the lowest limit. The highest values of overlap are observed for Skinner map combined with the Amber force field and the TCC scheme. The simulated spectra for this set of parameters together with experimental spectra are presented in Fig. 6. It is evident that there is room for improvement of the simulated spectra even for this simulation protocol. In Fig. 7, the spectra with the

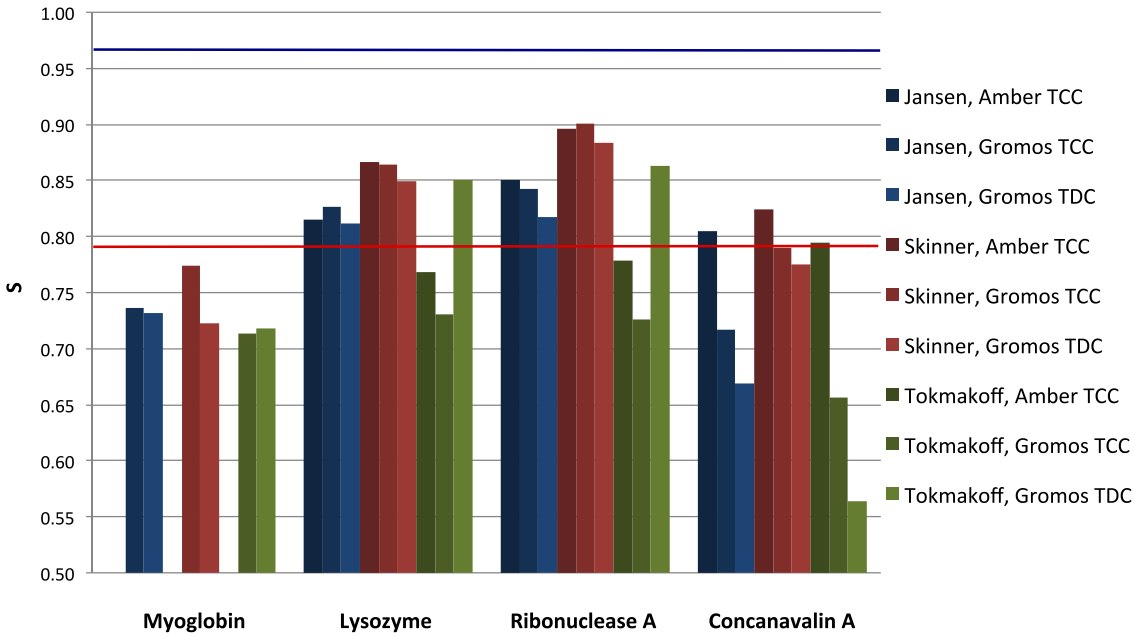

FIG. 5. Diagram of the spectral overlap between simulated 2DIR spectra and experimental ones for each of the four proteins. Upper limit (blue line) is the highest spectral overlap between experimental spectra (myoglobin and ribonuclease A). Lower limit (red line) is the lowest spectral overlap between experimental spectra (lysozyme and concanavalin A).

Amber force field and TCC for lysozyme are presented for the different electrostatic mappings. It is quite clear that the peak position is best reproduced with the Jansen map. The line shape for the Tokmakoff map is much too narrow and both the Jansen and Skinner maps underestimate the anti-diagonal linewidth.

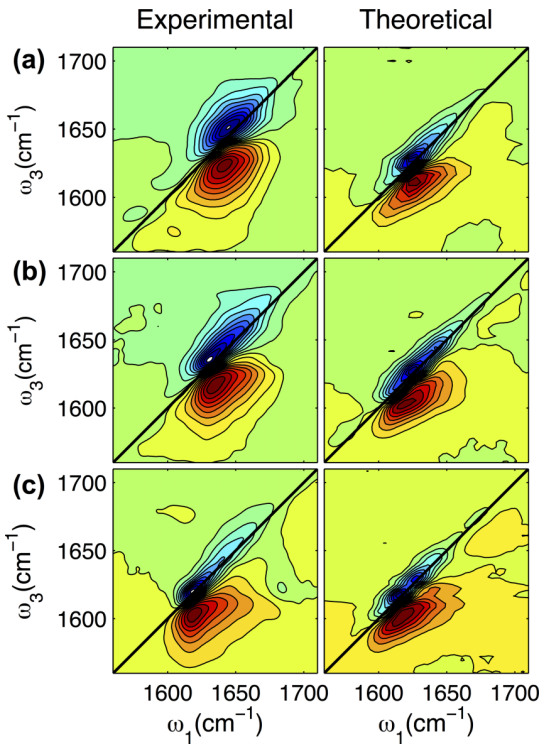

FIG. 6. The experimental 2DIR spectra of three of the four proteins compared to the theoretical spectra with the best performing model (Skinner map, Amber force field, and TCC): (a) lysozyme, (b) ribonuclease A, (c) concanavalin A. The contour lines were plotted equidistantly separated by 10% of the maximum value.

C. Discussion

We demonstrated that 2DIR spectroscopy is better for benchmarking spectral simulation protocols for the amide I band. As the full simulations of protein spectra are rather time consuming, we were limited to testing nine different combinations of parameters. Within this test set, even the best combination showed room for improvement and did not live up to the ideal situation allowing to discriminate between the spectra of proteins a similar as myoglobin and ribonuclease A. Among many simulation protocols developed, but not included in our test, a better method may, of course, exist.

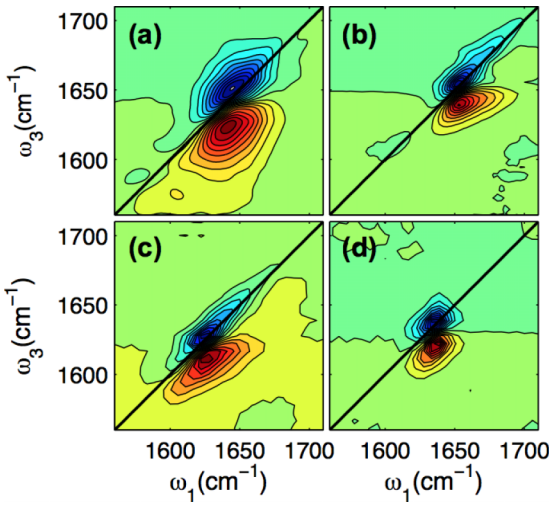

FIG. 7. The experimental 2DIR spectrum of lysozyme compared to the spectra simulated with the Amber force field, TCC, and three different electrostatic mapping schemes. (a) Experimental spectrum, (b) Jansen map, (c) Skinner map, (d) Tokmakoff map. The contour lines were plotted equidistantly separated by 10% of the maximum value.

The simulation protocols may be improved in many ways. In the present simulations, the side chain amide modes were not included. Here, we only tested one model for the nearest neighbour interactions. For the 2DIR spectra, we assumed a fixed anharmonicity of 16 cm^{-1} and a fixed lifetime of 1 ps during the coherence times. These parameters may also need further optimisation. The Tokmakoff map was originally developed for the CHARMM-27 force field[63] and the Skinner map for the GROMOS-53a6 force field.[60] The fact that we did not use those force fields here may, thus, explain at least the relatively poor performance of the Tokmakoff map in the present study. It is important to note that to eventually be able to simulate high quality spectra, one need both to have a correct description of the protein structures and dynamics and an accurate way to determine the vibrational Hamiltonian. Furthermore, the spectra were obtained from relatively short MD trajectories, thus relying on the assumption that the protein structures are relatively rigid and one structure dominates the configurational space in solution. In reality, one may need longer trajectories to account for the flexibility of the protein structures.

The choice of water model may certainly affect the interactions and spectra. In the present study, the simulations were performed exclusively using the SPC/E water model. As the AMBER and GROMOS force fields were parametrised with TIP3P (Transferable Intermolecular Potential 3P) and SPC (simple point charge), respectively, these force fields may appear as a more natural choice. However, our reason for using the SPC/E force field is that it predicts dynamics properties as the self-diffusion constant of water very well, whereas the SPC and TIP3P force fields predict too large diffusion constants (about twice too high for TIP3P).[116] The two-dimensional spectra are highly sensitive to the dynamics of the solvent, and therefore, we decided to consistently use the same water force field throughout the simulations. We recognise that it would be interesting to examine the effect of changing the water force field in the future.

The presented comparison with experimental data also poses some limitations and challenges. As already discussed, proteins as concanavalin A oligomerize or aggregate even at relatively low protein concentrations potentially leading to distortions of the spectra. Other experimental conditions may affect the spectra as well. These include effects of pH, temperature, ions in the buffer solutions, and pulse shapes of the used laser pulses.[2] Some of these effects may potentially be accounted for in the future.

We used the error in the frequency shift needed to maximize the spectral overlap and the spectral overlap as measures for the performance of the simulation protocols. One could think that the root mean square deviation between the simulated and experimental signals normalized with respect to the maximum peak height or peak volume would be a more natural measure, where the spectral overlap is equivalent to a root mean square deviation with a less intuitive normalization of $\sqrt{\sum_i I(\omega_i)^2}$. However, maximum peak heights are very sensitive to the signal at one single place in the spectra and the peak volume is not suited for 2DIR spectra, where the integrated volume of the peaks is close to zero. We, therefore, think that our choice provides more reliable results. Eventually, ideally one would like to use the overlap only without the introduction of peak shifts.

We foresee that the use of more accurate force fields including polarisability and multipole effects[117] may improve the simulation results in the future. This may change the balance between the mappings employed as it was already demonstrated that the Jansen map performed significantly better for NMA in chloroform, when such effects were included.[73] One can, thus, imagine that this mapping, which was not developed for a particular force field, but using fixed point charge environments, will eventually perform better than mappings developed for specific point charge force fields. Potentially, of course, new mappings may need to be developed instead for future and recently developed force fields that describe the protein structure and dynamics better than present day force fields.

IV. CONCLUSIONS AND PERSPECTIVES

Here, we demonstrated a novel benchmarking scheme for validating models for the amide I band of proteins. We compared the errors in peak positions and the spectral overlaps of the calculated FTIR and 2DIR spectra for four proteins and well-known structure and spectra. As one would expect, we found that 2DIR spectra provide a better discrimination between models than FTIR spectroscopy. From the nine different parameter combinations that we tried, the best one still leaves considerable room for future improvement. If the error in the peak position is systematic, one can easily compensate for this with a constant frequency shift. However, we found that in all the tested models the peak value errors were 10 cm^{-1} more to the blue for β-sheet dominated proteins as compared to α-helix dominated ones. The spectral overlaps provide a convenient measure of the ability of the different protocols to produce the correct line shapes. The presented benchmarking scheme should be a helpful tool to validate future simulation protocols and aid the development of better protein force fields, frequency mapping schemes, and coupling models. We foresee that developments in computational codes and the general increase in available computer power testing as more exhaustive combination of mappings and force fields will be within reach in a foreseeable future. Considering the popularity of the CHARMM-27 and OPLS-AA (Optimized Potentials for Liquid Simulations, All Atoms) force fields, these are obvious candidates for future benchmarking.

ACKNOWLEDGMENTS

A.S.B. acknowledges the support of the Erasmus Mundus programme of the European Union. The authors are grateful to Carlos Baiz and Andrei Tokmakoff for helpful discussions and for sharing their experimental data, which is now accessible at http://hdl.handle.net/1721.1/96430. Santanu Roy is kindly acknowledged for updating the AmideIMaps program to include all the employed mapping methods. We are grateful to Ana V. Cunha for providing useful suggestions to the manuscript.

[1]P. Hamm, M. H. Lim, and R. M. Hochstrasser, J. Phys. Chem. B **102**, 6123 (1998).
[2]P. Hamm and M. T. Zanni, *Concepts and Methods of 2D Infrared Spectroscopy* (Cambridge University Press, Cambridge, 2011).
[3]S. Woutersen and P. Hamm, J. Chem. Phys. **115**, 7737 (2001).
[4]J. Bredenbeck and P. Hamm, J. Chem. Phys. **119**, 1569 (2003).
[5]J. Bredenbeck, J. Helbing, J. R. Kumita, G. A. Woolley, and P. Hamm, Proc. Natl. Acad. Sci. U. S. A. **102**, 2379 (2005).
[6]P. Hamm, J. Helbing, and J. Bredenbeck, Annu. Rev. Phys. Chem. **59**, 291 (2008).
[7]A. Ghosh, J. Qiu, W. F. DeGrado, and R. M. Hochstrasser, Proc. Natl. Acad. Sci. U. S. A. **108**, 6115 (2011).
[8]C. Fang, J. Wang, A. K. Charnley, W. Barber-Armstrong, A. B. Smith III, S. M. Decatur, and R. M. Hochstrasser, Chem. Phys. Lett. **382**, 586 (2003).
[9]P. Mukherjee, A. T. Krummel, E. C. Fulmer, I. Kass, I. T. Arkin, and M. T. Zanni, J. Chem. Phys. **120**, 10215 (2004).
[10]D. B. Strasfeld, Y. L. Ling, S.-H. Shim, and M. T. Zanni, J. Am. Chem. Soc. **130**, 6698 (2008).
[11]J. Manor, P. Mukherjee, Y.-S. Lin, H. Leonov, J. L. Skinner, M. T. Zanni, and I. T. Arkin, Structure **17**, 247 (2009).
[12]A. M. Woys, Y.-S. Lin, A. S. Reddy, W. Xiong, J. J. de Pablo, J. L. Skinner, and M. T. Zanni, J. Am. Chem. Soc. **132**, 2832 (2010).
[13]S. D. Moran, A. M. Woys, L. E. Buchanan, E. Bixby, S. M. Decatur, and M. T. Zanni, Proc. Natl. Acad. Sci. U. S. A. **109**, 3329 (2012).
[14]H. S. Chung, Z. Ganim, K. C. Jones, and A. Tokmakoff, Proc. Natl. Acad. Sci. U. S. A. **104**, 14237 (2007).
[15]Z. Ganim and A. Tokmakoff, Biophys. J. **91**, 2636 (2006).
[16]Z. Ganim, H. S. Chung, A. W. Smith, L. DeFlores, K. C. Jones, and A. Tokmakoff, Acc. Chem. Res. **41**, 432 (2008).
[17]N. Demirdöven, C. M. Cheatum, H. S. Chung, M. Khalil, J. Knoester, and A. Tokmakoff, J. Am. Chem. Soc. **126**, 7981 (2004).
[18]Z. Ganim, K. C. Jones, and A. Tokmakoff, Phys. Chem. Chem. Phys. **12**, 3579 (2010).
[19]J. Lessing, S. Roy, M. Reppert, M. D. Baer, D. Marx, T. L. C. Jansen, J. Knoester, and A. Tokmakoff, J. Am. Chem. Soc. **134**, 5032 (2012).
[20]C. R. Baiz, Y.-S. Lin, C. S. Peng, K. A. Beauchamp, V. A. Voelz, V. S. Pande, and A. Tokmakoff, Biophys. J. **106**, 1359 (2014).
[21]C. R. Baiz, C. S. Peng, M. Reppert, K. C. Jones, and A. Tokmakoff, Analyst **137**, 1739 (2012).
[22]M. T. Zanni, N. H. Ge, Y. S. Kim, S. M. Decatur, and R. M. Hochstrasser, Biophys. J. **82**, 66 (2002).
[23]H. Maekawa, C. Toniolo, A. Moretto, Q. B. Broxterman, and N.-H. Ge, J. Phys. Chem. B **110**, 5834 (2006).
[24]N. Sengupta, H. Maekawa, W. Zhuang, C. Toniolo, S. Mukamel, D. J. Tobias, and N. H. Ge, J. Phys. Chem. B **113**, 12037 (2009).
[25]H. Meuzelaar, K. A. Marino, A. Huerta-Viga, M. R. Panman, L. E. J. Smeenk, A. J. Kettelarij, J. H. van Maarseveen, P. Timmerman, P. G. Bolhuis, and S. Woutersen, J. Phys. Chem. B **117**, 11490 (2013).
[26]H. Meuzelaar, M. Tros, A. Huerta-Viga, C. N. van Dijk, J. Vreede, and S. Woutersen, J. Phys. Chem. Lett. **5**, 900 (2014).
[27]L. Stryer, J. L. Tymoczko, and J. M. Berg, *Biochemistry*, 5th ed. (Freeman, New York, 2002).
[28]A. Barth, Biochim. Biophys. Acta **1767**, 1073 (2007).
[29]A. Barth, Prog. Biophys. Mol. Bio. **74**, 141 (2000).
[30]T. L. C. Jansen and J. Knoester, Biophys. J. **94**, 1818 (2008).
[31]H. Torii, Vib. Spectrosc. **42**, 140 (2006).
[32]T. L. C. Jansen and J. Knoester, J. Phys. Chem. B **110**, 22910 (2006).
[33]C. Liang and T. L. C. Jansen, J. Chem. Theory Comput. **8**, 1706 (2012).
[34]W. Zhuang, D. Abramavicius, T. Hayashi, and S. Mukamel, J. Phys. Chem. B **110**, 3362 (2006).
[35]T. L. C. Jansen and J. Knoester, Acc. Chem. Res. **42**, 1405 (2009).
[36]W. L. Jorgensen and J. Tirado-Rives, J. Am. Chem. Soc. **110**, 1657 (1988).
[37]K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, R. Dror, and D. Shaw, Proteins: Struct., Funct., Bioinf. **78**, 1950 (2010).
[38]U. Stocker and W. F. van Gunsteren, Proteins: Struct., Funct., Bioinf. **40**, 145 (2000).
[39]B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, J. Comput. Chem. **4**, 187 (1983).
[40]Y. Mu, D. S. Kosov, and G. Stock, J. Phys. Chem. B **107**, 5064 (2003).
[41]K. A. Beauchamp, Y.-S. Lin, R. Das, and V. S. Pande, J. Chem. Theory Comput. **8**, 1409 (2012).
[42]S. Gnanakaran and A. E. Garcia, J. Phys. Chem. B **107**, 12555 (2003).
[43]O. Guvench and A. D. Mackerell, *Molecular Modeling of Proteins* (Humana Press, 2008), Vol. 443, p. 63.
[44]N. G. Mirkin and S. Krimm, J. Mol. Struct.: THEOCHEM **377**, 219 (1996).
[45]P. Bour and T. A. Keiderling, J. Chem. Phys. **119**, 11253 (2003).
[46]J. H. Choi, S. Y. Ham, and M. Cho, J. Phys. Chem. B **107**, 9132 (2003).
[47]S. Ham, J. H. Kim, H. Lee, and M. H. Cho, J. Chem. Phys. **118**, 3491 (2003).
[48]S. Ham and M. Cho, J. Chem. Phys. **118**, 6915 (2003).
[49]K. Kwac and M. H. Cho, J. Chem. Phys. **119**, 2247 (2003).
[50]R. D. Gorbunov, D. S. Kosov, and G. Stock, J. Chem. Phys. **122**, 224904 (2005).
[51]T. M. Watson and J. D. Hirst, Mol. Phys. **103**, 1531 (2005).
[52]T. L. C. Jansen and J. Knoester, J. Chem. Phys. **124**, 044502 (2006).
[53]R. D. Gorbunov and G. Stock, Chem. Phys. Lett. **437**, 272 (2007).
[54]T. Hayashi and S. Mukamel, J. Phys. Chem. B **111**, 11032 (2007).
[55]R. Bloem, A. G. Dijkstra, T. L. C. Jansen, and J. Knoester, J. Chem. Phys. **129**, 055101 (2008).
[56]K. Cai, C. Han, and J. Wang, Phys. Chem. Chem. Phys. **11**, 9149 (2009).
[57]Y.-S. Lin, J. M. Shorb, P. Mukherjee, M. T. Zanni, and J. L. Skinner, J. Phys. Chem. B **113**, 592 (2009).
[58]J. Jeon and M. Cho, New J. Phys. **12**, 065001 (2010).
[59]S. Roy, J. Lessing, G. Meisl, Z. Ganim, A. Tokmakoff, J. Knoester, and T. L. C. Jansen, J. Chem. Phys. **135**, 234507 (2011).
[60]L. Wang, C. T. Middelton, M. T. Zanni, and J. L. Skinner, J. Phys. Chem. B **115**, 3713 (2011).
[61]A. Bastida, M. A. Soler, J. Zuniga, A. Requena, A. Kalstein, and S. Fernández-Alberti, J. Phys. Chem. B **116**, 2969 (2012).
[62]E.-L. Karjalainen, T. Ersmark, and A. Barth, J. Phys. Chem. B **116**, 4831 (2012).
[63]M. Reppert and A. Tokmakoff, J. Chem. Phys. **138**, 134116 (2013).
[64]E. Malolepsza and J. E. Straub, J. Phys. Chem. B **118**, 7848 (2014).
[65]R. D. Gorbunov, P. H. Nguyen, M. Kobus, and G. Stock, J. Chem. Phys. **126**, 054509 (2007).
[66]M. H. Farag, M. F. Ruiz-Lopez, A. Bastida, G. Monard, and F. Ingrosso, J. Phys. Chem. B **118**, 6186 (2014).
[67]F. Ingrosso, G. Monard, M. H. Farag, A. Bastida, and M. F. Ruiz-Lopez, J. Chem. Theory Comput. **7**, 1840 (2011).
[68]H. Torii, J. Phys. Chem. A **108**, 7272 (2004).
[69]H. Torii, J. Phys. Chem. Lett. **6**, 727 (2015).
[70]B. Blasiak, H. Lee, and M. Cho, J. Chem. Phys. **139**, 044111 (2013).
[71]B. Blasiak and M. H. Cho, J. Chem. Phys. **140**, 164107 (2014).
[72]J. H. Choi and M. H. Cho, J. Chem. Phys. **134**, 154513 (2011).
[73]T. L. C. Jansen, J. Phys. Chem. B **118**, 8162 (2014).
[74]T. L. C. Jansen, A. G. Dijkstra, T. M. Watson, J. D. Hirst, and J. Knoester, J. Chem. Phys. **125**, 044312 (2006).
[75]M. F. DeCamp, L. DeFlores, J. M. McCracken, A. Tokmakoff, K. Kwac, and M. Cho, J. Phys. Chem. B **109**, 11016 (2005).
[76]E. G. Buchanan, W. H. James, S. H. Choi, L. Guo, S. H. Gellman, C. W. Muller, and T. S. Zwier, J. Chem. Phys. **137**, 094301 (2012).
[77]J. Carr, A. Zabuga, S. Roy, T. Rizzo, and J. L. Skinner, J. Chem. Phys. **140**, 224111 (2014).
[78]S. Cha, S. Ham, and M. Cho, J. Chem. Phys. **117**, 740 (2002).
[79]H. Torii and M. Tasumi, J. Raman Spectrosc. **29**, 81 (1998).
[80]P. Hamm and S. Woutersen, Bull. Chem. Soc. Jpn. **75**, 985 (2002).
[81]S. Ham, S. Cha, J. H. Choi, and M. Cho, J. Chem. Phys. **119**, 1451 (2003).
[82]T. L. C. Jansen, A. G. Dijkstra, T. M. Watson, J. D. Hirst, and J. Knoester, J. Chem. Phys. **136**, 209901 (2012).
[83]S. Krimm and Y. Abe, Proc. Natl. Acad. Sci. U. S. A. **69**, 2788 (1972).
[84]H. Torii and M. Tasumi, J. Chem. Phys. **96**, 3379 (1992).
[85]A. M. Woys, A. M. Almeida, L. Wang, C.-C. Chiu, M. McGovern, J. J. de Pablo, J. L. Skinner, S. H. Gellman, and M. T. Zanni, J. Am. Chem. Soc. **134**, 19118 (2012).
[86]M. Reppert, A. R. Roy, and A. Tokmakoff, J. Chem. Phys. **142**, 125104 (2015).
[87]E. L. Hahn, Phys. Rev. **80**, 580 (1950).
[88]O. Golonzka, M. Khalil, N. Demirdöven, and A. Tokmakoff, J. Chem. Phys. **115**, 10814 (2001).
[89]J. F. Cahoon, K. R. Sawyer, J. P. Schlegel, and C. B. Harris, Science **319**, 1820 (2008).
[90]T. L. C. Jansen and J. Knoester, J. Chem. Phys. **127**, 234502 (2007).
[91]M. Ji and K. J. Gaffney, J. Chem. Phys. **134**, 0445516 (2011).
[92]C. Scheurer and T. Steinel, ChemPhysChem **8**, 503 (2007).
[93]J. Zheng, K. Kwak, J. B. Asbury, X. Chen, I. R. Piletic, and M. D. Fayer, Science **309**, 1338 (2005).
[94]S. Woutersen, Y. Mu, G. Stock, and P. Hamm, Chem. Phys. **266**, 137 (2001).
[95]N. Demirdöven, M. Khalil, and A. Tokmakoff, Phys. Rev. Lett. **89**, 237401 (2002).
[96]S. Roy, T. L. C. Jansen, and J. Knoester, Phys. Chem. Chem. Phys. **12**, 9347 (2010).

[97]S. T. Roberts, K. Ramasesha, and A. Tokmakoff, Acc. Chem. Res. **42**, 1239 (2009).
[98]J. T. King, C. R. Baiz, and K. J. Kubarych, J. Phys. Chem. A **114**, 10590 (2010).
[99]T. Takano, J. Mol. Biol. **110**, 569 (1977).
[100]P. Artymuik, C. Blake, D. Rice, and K. Wilson, Acta Crystallogr., Sect. B: Struct. Crystallogr. Cryst. Chem. **38**, 778 (1982).
[101]E. Chatani, R. Hayashi, H. Moriyama, and T. Ueki, Protein Sci. **11**, 72 (2002).
[102]A. Deacon, T. Gleichmann, A. Kalb, H. Price, J. Raftery, G. Bradbrook, J. Yariv, and J. Hellwell, J. Chem. Soc., Faraday Trans. **93**, 4305 (1997).
[103]E. Lindahl, B. Hess, and D. van der Spoel, J. Mol. Model. **7**, 306 (2001).
[104]N. Schmid, A. Eichenberger, A. Choutko, S. Riniker, M. Winger, A. Mark, and W. F. van Gunsteren, Eur. Biophys. J. **40**, 843 (2011).
[105]H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma, J. Phys. Chem. **91**, 6269 (1987).
[106]H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, J. Chem. Phys. **81**, 3684 (1984).
[107]W. G. Hoover, Phys. Rev. A **31**, 1695 (1985).
[108]M. Parrinello and A. Rahman, Phys. Rev. Lett. **45**, 1196 (1980).
[109]U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, J. Chem. Phys. **103**, 8577 (1995).
[110]See supplementary material at http://dx.doi.org/10.1063/1.4919716 for additional information including all obtained FTIR spectra, analysis using shifts to match the peak position, and analysis of the contribution to the FTIR spectra from different types of secondary structure.
[111]S. Krimm and J. Bandekar, Adv. Protein Chem. **38**, 181 (1986).
[112]R. M. Hochstrasser, Chem. Phys. **266**, 273 (2001).
[113]A. Barth and C. Zscherp, Q. Rev. Biophys. **35**, 369 (2002).
[114]J. Kruiger, C. P. van der Vegte, and T. L. C. Jansen, J. Chem. Phys. **142**, 054201 (2015).
[115]MATLAB and Statistics Toolbox Release, The MathWorks, Inc., Natick, Massachusetts, United States, 2013b. (2013).
[116]M. W. Mahoney and W. L. Jorgensen, J. Chem. Phys. **114**, 363 (2001).
[117]S. Jakobsen, T. Bereau, and M. Meuwly, J. Phys. Chem. B **119**, 3034 (2015).
[118]W. Humphrey, A. Dalke, and K. Schulten, J. Mol. Graphics **14**, 33 (1996).